L2 development in an intensive Study Abroad EAP context

Dr Parvaneh Tavakoli, Department of English anguage and Applied Linguistics, University of Reading

Accepted draft by System on 27-October-2017

L2 Development in an intensive Study Abroad EAP context

Abstract

The current study's main aim was to examine development of L2 proficiency over a short period of time by adopting an analytic framework that balances out the strengths and limitations of the existing Complexity, Accuracy, Lexis and Fluency (CALF) framework of linguistic measurement. CALF indices and discourse markers were used to analyse development of proficiency among participants on an intensive EAP course in a Study Abroad context. To investigate the differential rates of development in two different task types, gains in various aspects of proficiency were examined. The results suggest that some CALF measures adequately demonstrate L2 development over a one-month period. Discourse markers provide evidence of L2 development beyond CALF, and add a new dimension to investigating and measuring L2 development. The differences in L2 development indicated in monologic and dialogic performances imply that specific measures of analysis are more suitable to characterize development in different task types.

Key words: L2 development, CALF, communicative ability, discourse markers

Introduction

Since its emergence in the 1980s, task-based language teaching (TBLT) research has made a considerable contribution to the field of SLA by investigating the impact of task and task design on L2 performance, and by theorising the relationship between task performance and psycholinguistic processes involved in language production and comprehension (Ahmadian, 2012; Foster & Taavakoli, 2009; Kormos, 2006; Robinson,

2015; Skehan, 2015). Another major contribution of TBLT research to SLA has been the successful development of an analytic framework for operationalising and measuring syntactic complexity, accuracy, lexical complexity and fluency (CALF) in L2 performance. Research in this area has provided ample and robust evidence that CALF can usefully measure L2 performance (Ellis, 2009; Housen, Kuiken & Vedder, 2012; Skehan, 2009), and reliably predict L2 proficiency (de Jong, Stein, Florijn, Schoonen & Hulstijn, 2012; Revesz et al., 2016). While the findings of TBLT research have been central to our understanding of SLA processes and to providing a more in-depth awareness of processing demands associated with L2 learning, TBLT has been critiqued for predominantly focusing on task performance without adequately investing in research in L2 development (Lambert & Kormos, 2014; Pallotti, 2009). TBLT research has so far shed little light on L2 development as it has largely drawn on cross-sectional studies with data often collected under less ecologically valid conditions (Eckerth, 2008). Few systematic efforts have been made to examine the way L2 development progresses in instructional settings during a period of task-based instruction, to explore the differential effects of task on L2 development, or the kind of timescale within which development can be evident (for a full discussion see Ortega & Byrnes, 2008). Notwithstanding the significant contribution of TBLT to SLA, little attention has been paid to investigating the development of CALF over time in different task types and conditions frequently used in typical L2 classrooms. Similarly, examining the development of learner communicative ability has remained a relatively under-researched area in TBLT studies. These are particularly important as despite an increase in the number of students on short and intensive Study Abroad L2 courses (Hernandez, 2016), our knowledge of measuring L2 development in terms of CALF over a short period of time is limited. The study reported in this paper, therefore, sets out to help develop a better understanding of the extent to which L2 proficiency develops over

a short period of time in a Study Abroad context. The novel contribution of the current study lies in its developmental perspective, i.e. development of L2 over a period of time, and its adaptation of a framework that can provide a 'fuller' picture of L2 development than has been demonstrated in previous TBLT studies.

CALF measuring L2 performance and development

TBLT researchers have increasingly relied on measuring L2 proficiency by using CALF. Housen et al. (2012) argue that CALF have become standard measures employed by most researchers in TBLT, and are widely used in SLA research as performance descriptors, indicators of L2 proficiency, and evidence of L2 acquisition. Among several benefits of using the CALF framework, one can refer to its reliability in measuring linguistic performance (Pallotti, 2009; Skehan 1996, 2015), its potential for allowing researchers to employ a set of "more precise operationalisations of underlying constructs" (Skehan, 2001: 170), and its capacity to represent L2 ability in distinct constructs (Foster & Skehan, 1996; Michel et al.2007; Tavakoli & Foster, 2005). In addition, CALF are useful in helping understand the relationship between linguistic output and key cognitive processes underlying SLA, e.g. noticing, attention allocation and working memory (Robinson, 2011; Skehan, 2014), and lend themselves well to examination and interpretation of the processes involved in language production models, e.g. Levelt (1989) and Kellogg (1996). For example, use of CALF has enabled researchers (Kormos & Denes, 2004; Foster & Tavakoli, 2009; Tavakoli & Foster, 2008) to link a change in syntactic complexity of performance to variability in the cognitive demands that influence the Conceptualizer (where the preverbal message is generated), or to explain an increase in lexical diversity in terms of the processing needs of the Formulator (where the preverbal message is converted into a phonetic plan for speech), in Levelt's (1989) model of speech production.

Using CALF, however, is not without limitations. Some argue that there is not a linear relationship between CALF and communicative adequacy, i.e. L2 performance that is not highly accurate, complex or fluent, can be communicatively adequate if it conveys the intended message or achieves the task outcomes (de Jong et al. 2012; Kuiken & Vedder, 2012; Pallotti, 2009). The argument is based on the premise that it is possible to observe fluent and complex performance that does not fulfil the communicative needs of a given task. It has also been argued that although CALF are valid indicators of linguistic performance (Housen et al., 2012), they may fail to provide evidence of linguistic development. A disproportionately small number of studies have investigated L2 development by use of CALF. Gunnarsson (2012) and Ferrari, (2012), examining the development of a few learners' interlanguage in oral task performance over a period of three to four years, reported that although all L2 learners made some progress demonstrated in CALF, the rate of development varied considerably from task to task and learner to learner. Polat & Kim (2014), working with case-study data from an individual learner in a non-instructional setting over a period of a year, showed that the advanced L2 speaker's language development was visible in terms of lexical variety, with little improvement in syntactic complexity and almost no improvement in accuracy. These studies support use of CALF to detect L2 development over longer periods of time. There are a number of studies that have examined development of CALF in Study Abroad contexts over a period of three months (e.g. Leonard & Shea, 2017). However, they do not provide any evidence on whether CALF can trace L2 development over a shorter period of time, e.g. a month.

Finally, CALF are considered limited as they do not provide robust evidence of development in other aspects of performance, e.g. communicative and pragmatic abilities. Ortega (2003: 494) contends that development of learner language includes "syntactic complexification, but it also entails the development of discourse and

sociolinguistic repertoires that the language user can adapt appropriately to particular communication demands". In their seminal work evaluating CALF, Norris and Ortega (2009: 574) call for a more "organic practice" in measuring language performance and argue that in order to portray the complex, dynamic, and developmental nature of CALF phenomena, we need "multivariate, longitudinal, and descriptive accounts of constructs in L2 performance". A conclusion to make is that while CALF are effective and instrumental in measuring L2 performance, particularly in a linguistic sense, measures that can demonstrate L2 development in a more communicative sense are needed.

To provide a more overarching picture of L2 development, TBLT research also needs to investigate language performance in different task conditions. While studies in corrective feedback and negotiation of meaning (see Mackey & Gass, 2016, for further details) have mainly relied on dialogic forms of communication, dialogic tasks are less popular in TBLT research. The interest in using monologue for research purposes can perhaps be explained in the light of the degree of control in the linguistic units they elicit, and the ease and feasibility of data collection and coding. The intricacies involved in collecting and analysing dialogic data, on the other hand, has made dialogue an under-research task. Walsh (2013) among others argues that in an information-processing perspective to L2 acquisition, development has largely been investigated in monologic mode, with limited attention to measuring other task conditions that involve interaction between speakers.

By aiming to investigate L2 development beyond CALF, this study is an attempt to examine which other measure(s) can successfully reflect development of L2 communicative ability. Research has offered several perspectives on L2 development, but a point of contention in the debate of which perspective to choose is to identify measures that lend themselves effectively to the purpose of the study, i.e. examining L2

development in a communicative sense in a Study Abroad context. This will be discussed in the next section.

In search of measures of communicative ability

There is little disagreement among researchers that L2 proficiency ideally develops along both linguistic and communicative dimensions. In models of communicative language ability (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980), researchers have for a long time proposed that communicative ability is included as one of the components of linguistic ability. Bachman (1990) divides language competence to 'organisational competence', i.e. knowledge of linguistic units both at sentence and discourse levels, and 'pragmatic competence', i.e. knowledge of using language in socially appropriate ways (Bachman, 1990; Bachman & Palmer, 1996). While providing a detailed definition of pragmatic competence and discussing its relationship with *interactional competence* (Hall, 1991; Kramsh, 1986; Young, 2003) goes beyond the scope of the current paper, L2 pragmatics is simply considered as the study of how learners come to know "how-to-say-what-to-whom-when" successfully (Bardovi-Harlig, 2013: 68). Researchers investigating the development of L2 pragmatics often report that L2 learners in Study Abroad contexts develop competent discourse and pragmatic resources that facilitate successful communication and interaction (Bardovi-Harlig & Bastos, 2011; Barron, 2003; Hernandez, 2016; Matsumara, 2007). The underpinning assumption is that studying L2 in the target language community provides opportunities for learners to develop an awareness of the communicative skills needed to interact successfully, and allows them to link linguistic features of the L2 to the pragmatic needs of different speaking tasks (Kinginger & Blattner, 2008).

Lack of consensus among scholars on what to measure and how to measure it is perhaps one of the key challenges holding researchers back from researching development of L2 pragmatics. Bardovi-Harlig (2013: 76) argues that in the absence of global proficiency measures of pragmatics, SLA studies have often used "measures of development that are appropriate to the research questions posed and the research designs used to investigate them". Examples of local measures used in different studies include conventional expressions (Barron, 2003; Warga & Sch¨olmberger, 2007), conversational turn structure (Bardovi-Harlig & Salsbury, 2004), and semantic formulas (Shardakova, 2005). A language feature often associated with pragmatically successful communication is the use of discourse markers. Providing an operational definition for DMs as "intra-sentential and supra-sentential linguistic units which fulfil a largely non-propositional and connective function at the level of discourse", Fung and Carter (2007: 411) argue that DMs signal transitions in discourse, show relationships between different utterances, and promote interaction between speaker, hearer and message. Whether used in monologic or dialogic talk, Louwerse and Mitchell, (2003: 199) argue DMs "mark transition in discourse" and "facilitate the construction of a mental representation of the events described by the discourse". In dialogues, DMs play an important role as they act like "conversational glue that participants effectively use to hold the dialog together" (Louwerse & Mitchell, 2003). While DMs are typically more frequent in dialogues, research suggests that a range of non-interactional DMs are used in monologic talk to demonstrate cohesion and coherence and to structure and organize discourse (Louwerse & Mitchell, 2003; Pekarek Doehler & Berger, 2016). Two-word and longer DMs are formulaic in nature, often considered as a subset of formulaic sequences (see Conklin & Schmitt, 2008, for a full discussion).

DMs are "a pervasive and perceptually salient feature in colloquial English" (Lazaro & Garcio Mayo, 2012: 140) that play a fundamental role in making spoken interaction connected and coherent (Aijmer, 2002; Carter & McCarthy, 2006; Fung & Carter, 2007). Besides helping L2 speakers organize their discourse structure and achieve their communicative goals, researchers argue that DMs help learners make up for their limited linguistic resources by allowing them to establish relationships between different units of language and promoting communication between speaker and hearer (Hellermann & Vergun, 2007; Louwerse & Mitchell, 2003; Muller, 2004). Although corpus-linguistic studies (see Fung & Carter, 2007) testify a correlation between more proficient language use and more frequent use of DMs, use of DMs as a measure of L2 development has been rarely researched in TBLT studies. The rationale for studying DMs as a sign of L2 development in the current study is informed by the premise that development of pragmatic aspects of language use provides reliable evidence on the development of learner communicative ability (Fung & Cater, 2007; Hellermann & Vergun, 2007; Muller, 2004; Lazaro & Garcia Mayo, 2012; Schiffrin, 2001).

Research aims and questions

The current study is motivated by the question of whether CALF can measure L2 development in monologic and dialogic tasks over a short period of intensive instruction. In addition, the study is examining L2 development in terms of pragmatic use of language measured by the use of DMs. The following research questions (RQ) guide the study:

RQ1: To what extent do CALF portray L2 development over a one-month period of intensive TBLT instruction in a Study Abroad context at a university in the UK?

RQ2: To what extent does use of DMs help show L2 development over this period of time?

RQ3: To what extent does L2 development vary in different task conditions?

Methodology

Participants

The participants were 40 (25 male and 15 female) students studying English at a university in the UK. The sample size of the study, although not very large, is comparable to other longitudinal Study Abroad projects in which CALF are used to examine interlanguage development, e.g. 28 in Freed et al. (2004), 32 in Derwing et al. (2009), 39 in Leonard and Shea (2017) and 40 in Mora and Valls-Ferrer (2012). The participants had lived in the UK for up to a maximum of two months before data collection started, and had been enrolled on the course for five weeks when the first set of data were collected. The participants were young adults (mean age= 26.5) with a mix of different L1s including Arabic, Chinese, Russian, Portuguese, Kazakh, Thai and Korean. A standardized test of proficiency, TEEP (2014), measuring all the four skills had been used to place them on their course at B2 level of CEFR. Not all participants completed all tasks and tests during the data collection period, and hence the data reported here are from 37 and 35 participants performing monologues and dialogues respectively at the two times of data collection, i.e. Weeks 6 and 10 of the term.

The speaking component of the course adopted a task-based approach to teaching and learning, using a range of different communicative tasks related to the students' academic work at the university. Every week, the participants received 21 hours of instruction practicing all the four skills plus some coursework to complete. Given the Study Abroad context of the course, they were exposed to, and had various opportunities for using English for a range of purposes outside class both in and out of the university. However, the amount of L2 use outside classroom was not examined.

Design

The study had a 2 x 2 within-participants factorial design with two independent variables: Time (Time 1 versus Time 2) and Task (monologue versus dialogue). The dependent variables were CALF and use of Discourse Markers, as operationalised below. The data consisted of L2 learner task performances completed across the two times of the study, one month apart. At each time the learners performed a monologue individually, and a dialogue in a dyad with a partner. To control for any practice effect, a counterbalanced design was used in the order of performing the tasks.

Choosing a robust, valid and reliable research design is a key challenge in experimental studies. Although a within-participant design reduces the error variance associated with individual differences and allows for a more careful examination of individual's abilities, its limitation is that participation in one condition may influence performance in other conditions. A between-participant design is also limited as the individual differences among the participants can influence the outcomes. For example, given a Study Abroad experiment, differences between the participants in their language experience and use outside classroom are a major factor influencing their interlanguage development (Ranta & Meckelborg, 2013; Saito, 2015). As such, a within-participant design was considered more suitable for this study, and the use of two different but comparable task conditions was deemed justifiable.

To minimize the effects of task design on performance, a number of design features were carefully controlled for. Following previous research, the factors controlled for included *familiarity of information* (Bui, 2014), *familiarity with task type* (Bygate, 2001), *degree of contextual support* (Revesz, 2016), and *number of elements* in a task (de Jong & Vercelloti, 2015). It was also ensured that task instructions were similar in structure and the support provided. The two tasks were, however, different in that the

monologue involved describing past experiences, whereas the dialogue required discussion and persuasion (see Appendix 1). In order to have comparable data between the two points of data collection, the same task conditions were used, but to prevent a practice effect different topics were selected. The choice of the tasks and topics were discussed with the course teachers to ensure the tasks were suitable and the topics had not been used or practiced on the course before. Although research in SLA has provided ample evidence that repeating the same task, particularly when performed immediately or in short term intervals, promotes performance (de Jong & Perfetti, 2011; Thai et al., 2016), a repetition effect is not expected in this experiment as task topics varied and the two performances were one month apart.

CALF measures

Several researchers have highlighted a number of key problems in using CALF for evaluating L2 performance and development (Inoue, 2016; Lambert & Kormos, 2014; Norris & Ortega, 2009), and have underlined the importance of using CALF more carefully. Norris and Ortega (2009) and Inoue (2016) argue that a single measure of complexity may fail to portray a full picture of syntactic complexity, and suggest that using a battery of measures aiming for the four categories of a) subordination-based, b) length-based, c) coordination-based and d) phrasal complexity is necessary. In the current study, three of the four aspects of syntactic complexity set by Norris and Ortega (2009) are examined. They are ratio of subordination (category a), length of AS unit (Foster, Tonkyn & Wiggles worth, 2000) (Category b), and clause length (Category c). A measure of coordination was not deemed suitable as it develops at incipient proficiency levels (Norris & Ortega, 2009).

Based on the evidence about the robustness of global measures of accuracy (Ellis & Barkhuzein, 2005; Ong & Zhang, 2010), CALF studies have conventionally used

percentage of error-free clauses to represent accuracy. However, this measure has been criticised for failing to distinguish between errors with different degrees of seriousness. Arguing for a more finely tuned measure of accuracy, Foster and Wigglesworth (2016: 98) proposed a more systematic approach to measuring clause-level accuracy, i.e. a weighted clause ratio (WCR) measure, which "classifies errors at different levels" and distinguishes between "those that seriously impede communication, those that impair communication to some degree, and those that do not impair communication at all". Therefore, WCR was used to represent global accuracy in this study. Percentage of correct use of verbs (Ellis & Barkhuzein, 2005) was used as a local measure of accuracy.

Skehan (2003, 2009, 2014) argues that given its multifaceted nature, fluency should be measured in terms of speed, breakdown and repair. Recent research (Kahng, 2014; Tavakoli, 2016; Tavakoli, Campbell & McCormack, 2016; Witton-Davies, 2014) suggests that measures of length and speed of speech, and frequency and location of pauses are reliable measures that distinguish fluent from disfluent speech. The fluency measures employed in this study are mean length of run, speech rate, number of silent pauses clause internal and clause external, and a composite repair measure. The composite repair measure included repetitions, hesitations, reformulations and false starts. Although there is emerging evidence in the literature (de Jong, et al., 2012; Huensch & Tracey –Ventura, 2016) to suggest L2 fluency is related to L1 fluency and L1 background, given the multilingual population of the participants in this study it was not possible to control for L1 background or L1 fluency behaviour. The temporal aspects of fluency were calculated by use of PRAAT (Boersma & Weenik, 2007) and for 60 seconds of each participant's performance per task.

To demonstrate development in lexical diversity, i.e. "range of different words in a text" (McCarthy & Jarvis 2010: 381), measures of D (Malvern & Richards, 2002) and textual lexical diversity (MTLD) that are reported to be least affected by text length

were used (Graesser, McNamara, & Kulikowich, 2011; Graesser, McNamara, Louwerse, & Cai, 2004). Coh-Metrix (Graesser, McNamara, & Louwerse, 2003) was used to calculate D and MTLD. To ensure reliability of the data coding and analysis, 20 percent of the coded data, 5 percent of each task at each time, was randomly double blind coded by an experienced researcher and an inter-rater correlations of .88 to .95 was obtained for different CALF measures. For speed and breakdown fluency measures calculated in PRAAT, 20% of the data was coded by the same researcher, and an intra-rater correlation of .96 was obtained.

Discourse markers

Following from Schiffrin (1987) and Louwerse and Mitchell (2003), a linguistic unit is considered a DM if it is a sequentially dependent element that supports units of talk, marks a structural boundary, i.e. starts a new structural unit, and operates at both a global and local discourse level. Unlike Louwerse and Mitchell (2003), the criterion of prosodic contours was considered inappropriate as the data came from L2 learners who had not yet fully mastered the phonological patterns of spoken English and therefore were less likely to use prosodic contours consistently and correctly to mark their discourse.

The analysis of DMs followed a three-step procedure. First, adopting the criteria discussed above all one-word DMs were identified in the transcripts. Then, two-word and three-word and longer DMs were identified and coded. Following from Fraser's (2015: 51) analytic scheme, combined DMs such as *well obviously* and *but instead* were coded as two-word DMs. Finally, in order to identify common patterns of use of DMs and the extent of development, the data were examined qualitatively. The qualitative analysis examined DMs in three respects:

a) structural accuracy (whether two-word and longer DMs were correct formulaically)
b) structural complexity (whether some DMs were more complex in structure)
c) communicative efficiency (whether DMs were appropriately used for the communicative needs of the discourse, e.g. 'on the other hand' was used to show a contrast, and 'sorry for interruption' was used to interrupt the interlocutor)

Structural accuracy and communicative efficiency were subjectively rated by marking two-word, three-word and longer DMs as either correct or incorrect, and as communicatively effective or not effective. If needed, structural accuracy was checked against BNC, and communicative efficiency was checked by listening to the audio-recording of the data. As for analyzing structural complexity of DMs, research in this area (Aijmer, 2002; Fung & Carter, 2007; Louwerse & Mitchell, 2003) suggests that DMs are by nature largely non-conceptual or propositional, and as such their complexity cannot be measured in the same way as syntactic complexity.

Complexity of DMs was examined in terms of length and sophistication of their lexical components. As for length, it was assumed that acquisition of many, but not all, one word DMs occur before acquisition of two and three word DMs, e.g. use of *but* emerges before *on the other hand*. It is possible to argue that use of some one word DMs, e.g. *hence*, emerges after the development of some multi-word DMS, e.g. *on the other hand*. However, this aspect of complexity relates to lexical sophistication which will further be examined in this framework. For sophistication, it is plausible to assume that more frequent DMs are lexically less complex, e.g. *but* is less complex than *nevertheless*. Based on these assumptions, number of words in a DM was used to assess structural complexity in terms of length, and the 1K and 2K frequency lists (Vocab Profile, Cobb, 2015) were used to analyse DMs in terms of lexical sophistication. Although structural

complexity can also be reflected in the use of clausal or sub-clausal units in longer DMs, this measure was not used as it overlaps with the measures of syntactic complexity used for CALF analysis. Neither did the analysis categorize the different functions of DMs since one DM can sometimes be used for different functions (Fung & Carter, 2007). A researcher with expertise in discourse analysis second rated the coding of 20% of the data which led to a .92 kappa coefficient.

Analysis and results

The analysis and results section is presented according to Larson-Hall and Plonsky's (2015) recommendations. The descriptive statistics is provided in Table 1 for all the analytic measures used in the study in both task conditions and the two times of data collection. In addition to means and standard deviations, gains in these measures across time are provided to clarify the extent of development in each task condition. As can be seen in Table 1, the descriptive statistics (for group means) shows positive gains in most aspects of proficiency over the short period of time of the experiment. The only measures that show little or no change are clause-internal and clause-external pauses.

Insert Table 1 here

Further analyses were run to investigate whether these gains reached a statistically meaningful level. First, a repeated-measures multivariate analysis of variance (MANOVA) was run to investigate whether there were statistically significant differences in learners' L2 proficiency at the two times of data collection and between the two different tasks. Drawing on previous literature in this area (Foster & Skehan, 1996; Michel et al.2007; Tavakoli & Skehan, 2005), four measures, one from each category of CALF analysis (i.e. ratio of subordination, WCR, D and speech rate) which

are reported to represent CALF consistently, were selected from the total of 14 measures used in the analysis. The reduction from 14 to four measures was to satisfy the MANOVA requirement of having an acceptable ratio of dependent variables to the number of cases in each cell (minimum 5 cases for each variable according to Tabachnick & Fidell, 1996). The independent variables were Time and Task. All the preliminary assumptions of normality and linearity were checked to ensure no violations were observed. The non-significant results obtained for Kolmogorov-Smirnov tests confirmed the normality of the distribution of the data. The analysis confirmed the overall effect of Time, Task and the interaction between the two on dependent variables, indicating three statistically significant differences with noticeable effect sizes: one for Time (Wilks' Lambda = .286; F = 19.35, p = .000; η2 =.71), one for Task (Wilks' Lambda = .435; F = 10.06, p = .000; η2 =.56), and one for the interaction between Time and Task (Wilks' Lambda =.707; F = 3.21, p = .02; η2 =.29). These results allowed for further detailed analyses to examine the extent of improvement across time and between tasks.

Univariate analyses were run to identify the effects of time and task condition on performance. To avoid running an increased risk of Type 1 error, i.e. the risk of having some spurious alpha levels, a Bonferroni-adjusted alpha level of .025 was used. Cohen d effect sizes were calculated. Effect sizes are particularly important in experimental research as they allow researchers to go beyond the level of significance to show the magnitude of the difference between the groups. Plonsky and Oswald (2014) argue that "effect sizes are best understood when interpreted within a particular discipline or domain" (p. 878), and suggest d values of .40 as small, .70 as medium, and 1.00 as large for between-group means, and d values of .60 as small, 1.00 as medium, and 1.40 as large for within-group comparisons in applied linguistics studies.

RQ1: The results of the univariate analyses comparing gains in CALF from Time 1 to Time 2 indicated that only some of the gains reached a statistically significant level.

All the significant differences are highlighted in Table 1 above. For syntactic complexity, a significant difference was observed only in dialogue for ratio of subordination (t = 3.85, p < 000, d = .68) and length of AS unit (t = 2.43, p < 02, d = 0.41), with length of clause failing to reach a statistically significant level. For accuracy, while at Time 2 the WCR measure showed a statistically higher ratio of accuracy in monologue (t = 3.19, p < 004, d = .47), the changes were not significant in dialogue. As for correct use of verbs, the improvement failed to reach a statistically meaningful level in monologue (t = .44, p < .66), but in dialogue performance was statistically more accurate (t = 2.59, p < 01, d = 0.51). For fluency measures, the gains in mean length of run in monologue (t = 2.76, p < 009, d = 0.47) and dialogue (t = 3.91, p < 000, d = 0.43), and speech rate in monologue (t = 4.15, p < 000, d = 0.72) and dialogue (t = 3.68, p < 000, d = 0.67) reached a statistically significant level. Although repair fluency improved in time across both tasks, it failed to reach a significant level. Interestingly, clause-internal and clause-external pauses did not show much change across time in monologue or dialogue. In terms of lexical diversity, while the results showed little improvement in monologue, the gains reached a statistically meaningful level in dialogue for D (t = 3.03, p < 005, d = .62) and MTLD (t = 4.41, p < 000, d = .95). It is necessary to note that some of the obtained effect sizes are considered small, according to Plonsky and Oswald's (2014) guidelines. However, the effect size for speech rate and lexical diversity were larger than the others. In comparison, the effect sizes are larger than those obtained by Mora and Valls-Ferrer (2012), i.e. .024 to .51, indicating a more robust evidence of development.

RQ2 asked whether the use of DMs can demonstrate L2 development over this short period of time. The descriptive statistics (see Table 1) indicated that over time learners used more DMs (total number, two-word and longer units). The results of univariate t-tests showed that development in the use of DMs over time was significant for total number of DMs in monologue (t = 3.02, p < 005, d = .78), two-word DMs in both

monologue (t = 3.91, p < 000, d = .97) and dialogue (t = 3.05, p < 004, d = .68), and longer DMs in both monologue (t = 3.16, p < 003, d = .70) and dialogue (t = 4.57, p < 000, d = .96), all with small to medium effects sizes (Plonsky & Oswald, 2014). The largest effect sizes, i.e. 0.96 and 0.97, were observed for use of two-word and longer DMs in dialogue, suggesting that a considerable amount of variance in the use of DMs over time can be explained in the light of L2 development in dialogic tasks.

Table 2 shows the total number of DMs used by the participants in different tasks and across time, and provides examples from the data for qualitative illustration of the statistical patterns discussed above.

Insert Table 2 here.

As shown in Table 2, the participants used more DMs across both tasks at Time 2 of the data collection. This pattern of increase was observed for total number, two-word, and three-word and longer DMs, but not in one-word DMs. The qualitative analysis of DMs showed a few important patterns. Firstly, the most frequent one-word DMs at Times 1 and 2 were *and, because, so,* and *but* respectively. The analysis showed that many one-word DMs were replaced with two-word and longer DMs at Time 2 (e.g. *first of all* instead of *first*). This often involved combining one word DMs with other DMs, e.g. *and then*, *but actually*, which explains the decrease in the number of one-word discourse markers at Time 2. Second, in terms of structural accuracy, the analysis showed that at both times the learners used a large majority of the two-word DMs correctly, but there were a few deviations from the norm in the use of longer DMs (e.g. *let me pick you up* ***with*** *this, sorry for* ***interrupt*** *you*). These errors were all longer than three words and contained clausal or sub-clausal units.

As for communicative efficiency, a large majority of the DMs (88% at Time 1 and 93% at Time 2) were used communicatively efficiently (e.g. using *but actually* to disagree with the partner). Finally, for structural complexity, the analysis showed that the participants used longer DMs more frequently at time 2 particularly in dialogue (see the significant differences in two-word and longer DMs above). Examples of more complex DMs include *I know what you mean*, *I see where you are coming from*, and *actually you're right*. Interestingly, no statistically significant differences were observed with regard to the lexical sophistication of one-word DMs at the two times of study when DMs were checked against 1K and 2K frequency lists (Vocab Profile, Cobb, 2015). The t-tests showed no statistical differences between the means of 1K DMs ($p < .367$) or 2K DMs ($p < .411$) in monologue, and 1K DMs ($p < .167$) or 2K DMs ($p < .121$) in dialogue at the two times of the study.

RQ3 aimed at examining the extent to which L2 development varied in different task conditions, i.e. descriptive monologic versus persuasive dialogic. To answer this question, the results of the univariate analyses examining the effects of task condition are presented here. In terms of syntactic complexity, performance in dialogue was statistically more complex in terms of ratio of subordination ($t = 6.70$, $p < 000$, $d = 1.74$), and mean length of AS units ($t = 7.29$, $p < 000$, $d = 1.76$), both with large effect sizes. Although length of clause was longer in monologue, performance in dialogue was overall more syntactically complex in terms of subordination and length of AS units. Performance in dialogue was more fluent in terms of speech rate ($t = 5.47$, $p < 000$, $d = 1.25$) and repair measures ($t = 2.80$, $p < 008$, $d = 1.74$). There were no statistically significant differences between the two task conditions in either clause-internal or clause-external pauses. As for lexical diversity in D and MTLD, there were no large differences between the two conditions. The results of the univariate analysis comparing DMs in monologic and dialogic task performance showed that learners used more DMs

in dialogue than in monologue, reaching statistically significant levels with large effect sizes (Plonsky & Oswald, 2014). The comparisons that reached a significant level included the total number of DMs (t = 6.94, p < 001, d = 1.57), two-word DMs (t = 7.62, p < 000, d = 1.17), and longer DMS (t = 5.43, p < 000, d = 1.16). Interestingly, these effect sizes are the largest obtained in the study. The findings of the study with regard to the RQs will be discussed in the next section.

Figure 1 below shows the statistically significant differences in the measures between monologic and dialogic task performance at Time 2 of the study.

Insert Figure 1 here.

Discussion

The study set out to investigate development of L2 ability in an intensive course of EAP instruction in a Study Abroad context. This context is particularly important as thousands of students from different L1 backgrounds join these courses in international universities each year to receive intensive L2 training and to develop a certain amount of L2 proficiency usually over a short period of time. The study was also motivated by the question of what measures can help portray the communicative aspects of L2 development in this particular context.

The results indicated that a number of CALF measures showed L2 development over a period as short as a month. For syntactic complexity, the analysis indicated a statistically significant increase in ratio of subordination and mean length of AS unit in dialogue, but mean length of clause failed to show a significant change in the learners' performance. The statistically meaningful gains in the two measures of complexity in dialogue, but not in monologue, highlights the importance of choosing complexity measures more carefully as different measures may tap into different aspects of complexity and may be more suitable for different task conditions (Inoue, 2016; Norris &

Ortega, 2009). As for accuracy, the significant results for WCR in monologue and percentage of correct use of verbs in dialogue show the differences in measuring development of accuracy in different tasks, and suggest that learners may prioritise different aspects of accuracy in different task conditions. This finding also suggests that a single accuracy measure is insufficient to provide a full picture of development of accuracy. The results of accuracy and complexity combined suggest that in monologue the participants were in better control of their language at a clause level, while performance in the dialogue was more syntactically complex with a more accurate use of verbs. The more accurate language in monologue at clause level may be explained in the light of more frequent pauses the learners made in their monologic performance. Leonard and Shea (2017) report that although CALF measures develop over a period of three months Study Abroad experience, learners with higher linguistic knowledge and processing ability gained more in accuracy and complexity.

While pauses are essential during speech production, whether in L1 or L2, SLA researchers (Kormos, 2006; Skehan, 2009, 2014) argue that pauses are central to monitoring process, especially for less advanced learners (Levelt, 1989).

In terms of fluency, the results for fluency measures indicate that speech rate and mean length of run show development in both task conditions. This result is in line with previous findings that suggest speed fluency develops quickly in Study Abroad contexts (Mora & Valls-Ferrer, 2012; Tavakoli, et al., 2016). Kormos and Denes (2004) report that speed fluency measures highly correlate with expert norms of proficiency, and influence listener perceptions of proficient language use. The limited change in the participants' pausing behaviour observed here can partly be explained in the light of the short period of the intervention, though this is different from Vercelloti (2015) who reported significant improvement in learner pausing behaviour over a period of a month. As discussed above, pauses are generally known as monitoring tools during the process

of speech production (Kahng, 2014; Kormos, 2006; Skehan, 2014): clause-external pauses help speakers with *Conceptualisation* and clause-internal pauses are useful for *Formulation* of the speech production process (Kormos, 2006; Levelt, 1989). Given that a reduced number of pauses can indicate automaticity and efficiency of speech production (DeKeyser, 2007; Lambert & Kormos, 2014), it seems plausible to expect the pausing behaviour to develop over a longer period of exposure, instruction and practice. As regards repair measures, while there were fewer repairs over time, the gain did not reach a statistically significant level. The finding is in line with previous research (Kormos, 2006; Tavakoli, et al., 2016) that suggest repair measures are slow to develop, and may be constrained to some extent as a function of L1 speaking style (de Jong et al, 2013).

For lexical complexity, while developing a larger repertoire of lexical items over a short period of time may be difficult for most L2 learners, the results suggest that the learners used a more varied set of lexical items over time when performing a dialogue. The gains in lexical diversity (both D and MTLD) in dialogic performance can partly be explained in the light of the differences between the two tasks in that the dialogue was less controlled in terms of the linguistic units it elicited. In line with Michel's (2011) explanation, it is possible to speculate that in dialogue the exchanges between the two participants allow them to borrow lexis from each other and therefore adding to the variety of the lexis each person is using.

Another aim of the study was to engage with the argument of 'more complex does not necessarily mean communicatively more successful' (Ortega, 2003: 494). To expand measurement beyond CALF and to examine L2 development from a more communicative perspective, the study employed DMs as a measure of communicative ability that tap into pragmatic repertoires learners acquire during L2 development. The results suggest that DMs can supplement the L2 development profile by offering a

different perspective to L2 ability. The statistically significant differences, as well as medium to large effect sizes, between the use of DMs in Time 1 and Time 2 suggest that as part of the L2 development process the learners used more two-word, three-word and longer DMs to organize their discourse. Although the analysis of lexical sophistication showed no significant differences between the use of one-word DMs in Time 1 and Time 2, the qualitative analysis suggested that DMs were structurally more complex at Time 2 with an increase in their length. Frequent examples of complex DMs in Time 2, e.g. *sorry for interrupting you* and *can I just come in*, which were rarely seen in Time 1, provide evidence of the development of learners' pragmatic repertoires that facilitate successful communication.

From a psycholinguistic perspective, since longer DMs are mostly formulaic in nature (e.g. first of all, as a matter of fact), retrieving and processing them is easier than processing multiple individual linguistic units (Schmitt, 2000), hence facilitating the L2 production process. The interesting results obtained for DMs as a measure of development of L2 pragmatics opens up both a new perspective for understanding, operationalising and measuring L2 development and a new platform for engaging in debates on what may supplement CALF framework. Developing an analysis framework to examine DMs to demonstrate development of L2 communicative ability was a challenge this study faced. The framework proposed here, however, should be considered as an initial attempt to be subjected to further research and scrutiny.

The final aim of the study was to compare L2 development in monologic and dialogic task performance. Before discussing the findings, it is important to note that given the design of the study, it was not possible to use exactly the same tasks in different modes. As such, some of the findings can be attributed to the inherent differences between the two task conditions. Previous research (Michel, 2011; Witton-Davies, 2014) has reported several differences between monologic and dialogic task

performance, but not much is known about the extent to which measuring L2 development can be mediated by task condition. The results suggest that development of L2 ability may be reflected differently in different task conditions, e.g. persuasive requirement of the dialogue invites more subordination, longer AS units and more accurate use of verbs, whereas the descriptive nature of the monologue is associated with longer and more accurate clauses. Such findings highlight the importance of using different tasks when measuring development.

From a pedagogic perspective, the findings are reassuring for teachers as the results suggest the effects of instruction combined with the Study Abroad context conditions can provide L2 learners with rich opportunities for development even over a short period. Whether the effects of instruction, the conditions of Study Abroad, or the psychological and cognitive factors play a more important role in this development requires further research.

It is imperative to note that the findings of the study should be interpreted with care and caution as they are based on a small-scale study conducted over a short period of a month. A larger-scale study with participants of different proficiency levels, different task types and in longer terms can undoubtedly shed more light on the development of CALF measures over time.

References

Ahmadian, M. 2012. The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly, 46*(1), 165-175.

Aijmer, K. 2002. *English discourse particles: Evidence from a corpus* (Vol. 10). Amsterdam: John Benjamins Publishing.

Bachman, L. 1990. *Fundamental considerations in language testing*. Oxford: Oxford Univrsity Press.

Bachman, L., and Palmer, A. 1996. *Language testing in practice: designing and developing useful language tests*. Oxfrod: Oxford University Press.

Bardovi-Harlig, K., and Bastos, M.-T. 2011. Proficiency, length of stay, and intensity of interaction and the acquisition of conventional expressions in L2 pragmatics. *Intercultural Pragmatics, 8*(3), 347-384.

Bardovi-Harlig, K. 2013. Developing L2 pragmatics. *Language Learning, 63*(s1), 68-86.
Barron, A. 2003. *Acquisition in interlanguage pragmatics: Learning how to do things with words in a study abroad context* (Vol. 108). Amsterdam: John Benjamins.
Boersma, P., and Weenink, D. 2007. Praat version 4.5.01. Computer software, downloaded from *http://www.fon.hum.uva.nl/praat/*.
Bygate, M. 2001. Effects of task repetition on the structure and control of language. In M. Bygate, P. Skehan and M. Swains (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 23-48). Harlow, Essex: Longman.
Canale, M., and Swain, M. 1980. Theoretical bases of com-municative approaches to second language teaching and testing. *Applied linguistics, 1*(1), 1-47.
Carter, R., and McCarthy, M. 2006. *Cambridge grammar of English: A comprehensive guide to spoken and written grammar and usage*. Cambridge: Cambridge UP.
Clark, H., and Fox Tree, J. 2002. Using uh and um in spontaneous speaking. *Cognition, 84*(1), 73-111.
Cobb,T. 2015. Web Vocabprofile [accessed December 2015 from http://www.lextutor.ca/vp/
Cohen, J. 1988. *Statistical power analysis for the behavioral sciences (2nd ed)*. New Jersy: Lawrence Erlbaum Associates.
Coxhead, A. 2000. A new academic word list. *TESOL Quarterly*. 34(2), 213–238.
Conklin, K. & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72-89.
De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., and Hulstijn, J. H. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(05), 893-916.
De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., and Hulstijn, J. H. 2012. Facets of speaking proficiency. *Studies in Second Language Acquisition*, 34(01), 5-34.
De Ruiter, J. P., Mitterer, H., and Enfield, N. J. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 515-535.
DeKeyser, R. 2007. *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge: Cambridge University Press.
Dewaele, J.-M. 1996. How to measure formality of speech? A Model of Synchronic Variation. *Approaches to second language acquisition. Jyväskylä Cross-Language Studies, 17*, 119-133.
Eckerth, J. 2009. Negotiated interaction in the L2 classroom. *Language Teaching*, 42(01), 109-130.
Ellis, R. 2009. The differential effects of three types of task planning on the fluency, complexity and accuracy in L2 oral production. *Applied Linguistics, 30*(4), 474-509.
Ellis, R., and Barkhuizen, G. 2005. *Analyzing learner language*. Oxford: Oxford University Press.
Ferrari, S. 2012. A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken and I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277-297). Amsterdam: John Benjamins.
Foster, P., and Skehan, P. 1996. The influence of planning and task type onsecond language performances. Studies in Second Language Acquisition, 18, 299-323.
Foster, P. & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity. *Language Learning*, 59(4): 866-896.

Foster, P. and Wigglesworth, G. 2016. Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98-116.
Foster, P., Tonkyn, A., and Wigglesworth, G. 2000. Measuring spoken language. *Applied Linguistics, 21*(3), 354-375.
Fraser, B. 2015. Classroom learning environments. *Encyclopedia of Science Education* (pp. 154-157). Netherlands: Springer.
Fung, L., and Carter, R. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied linguistics, 28*(3), 410-439.
Gilabert, R. 2007. Effects of manipulating task complexity on self-repairs during L2 production. IRAL, 45, 215-240.
Graesser, A., McNamara, D., and Kulikowich, J. 2011. Coh-Metrix providing multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.
Graesser, A., McNamara, D., and Louwerse, M. 2003. What do readers need to learn in order to process coherence relations in narrative and expository text. In C. Snow, A. Sweet and G. Press (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. 2004. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, instruments, and computers, 36*(2), 193-202.
Gunnarsson, C. 2012. The development of complexity, accuracy and fluency in the written production of L2 French. In A. Housen, F. Kuiken and I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA* (pp. 247-276). Amsterdam: John Benjamins.
Hall, J. K. 1999. A prosaics of interaction: The development of interactional competence in another language. In E. Hinkel (Ed.), *Culture in second language teaching and learning* (pp. 137-151). New York: Cambridge University Press.
Halliday, M., and Hasan, R. 1976. *Cohesion in English*. London: Longman.
Hernandez, T. 2016. Acquisition of L2 Spanish requests in short-time study abroad. Study Abroad *Research in Second Language Acquisition and International Education, 1*(2), 186-216.
Hellermann, J., and Vergun, A. 2007. Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics, 39*(1), 157-179.
Huensch, A. and Tracey –Ventura, N. 2016. Understanding second language fluency behaviour: The effects of individual differences in first language fluency, cross-linguistic differences and proficiency over time. *Applied Psycholinguistics*, online version.
Housen, A., Kuiken, F., and Vedder, I. 2012. Complexity, accuracy and fluency. In A. Housen, F. Kuiken and I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (Vol. 32, pp. 1-20). Amsterdam: John Benjamins.
Kahng, J. 2014. Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809-854.
Kellogg, R. T. 1996. A model of working memory in writing. In C. M. Levy and S. Randsell (Eds.), *The science of writing: Theories, methods, individual differences and applications* (pp. 57-72). Mahwa, NJ: Erlbaum.
Kinginger, C., and Blattner, G. 2008. Histories of engagement and sociolinguistic awareness in study abroad. In L. Ortega and B. H. (Eds.), *The longitudinal study of advanced L2 capabilities* (pp. 223–246). New York: Routledge.
Kormos, J. 2006. *Speech production and second language acquisition*. London: Routledge.

Kormos, J., and Denes, M. 2004. Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(1), 145-164.
Kramsch, C. 1986. From Language Proficiency to Interactional Competence. *Modern Language Journal*, 70(4), 366-72.
Kuiken, F., and Vedder, I. 2012. Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken and I. Vedder (Eds.), *Dimensions of L2 Performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 143-169). Amsterdam: John Benjamins.
Lambert, C., and Kormos, J. 2014. Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics*, 607–614.
Larsen-Freeman, D. 2006. The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied linguistics, 27*(4), 590-619.
Larsen-Freeman, D. 2009. Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579-589.
Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*(1), 127-159.
Lázaro, A., and Garcia Mayo, M. d. P. 2012. L1 use and morphosyntactic development in the oral production of EFL learners in a CLIL context. *International Review of Applied Linguistics in Language Teaching, 50*(2), 135-160.
Leonard, K. & Shea, C. (2017). L2 speaking development during study abroad: Fluency, accuracy and complexity, and underlying cognitive factors. *The Modern Language Journal, 101*(1), 179-194.
Levelt, W. 1989. *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
Louwerse, M. M., and Mitchell, H. H. 2003. Toward a taxonomy of a set of discourse markers in dialog: A theoretical and computational linguistic account. *Discourse processes, 35*(3), 199-239.
Mackey, A. & Gass, S. 2016. *Second language research: Methodology and design.* London: Routledge.
Malvern, D., & Richards, B. 2002. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, 19(1), 85–104.
Matsumura, S. 2007. Exploring the aftereffects of study abroad on interlanguage pragmatic development. *Intercultural Pragmatics, 4*(2), 167-192.
McCarthy, P. M., and Jarvis, S. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods, 42*(2), 381-392.
Michel, M., Kuiken, F., and Vedder, I. 2007. The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 241-259.
Müller, S. 2004. 'Well you know that type of person': Functions of well in the speech of American and German students. *Journal of Pragmatics,36(4)*, 1157-82.
Myles, F. 2008. Investigating learner language development with electronic longitudinal corpora. In L. Ortega and H. Byrne (Eds.), *The Longitudinal Study of Advanced L2 Capacities* (pp. 58–72). London: Routledge.
Norris, J., and Ortega, L. 2009. Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-577.

Ong, J., and Zhang, L. 2010. Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19*(2), 218-233.
Ortega, L. 2003. Syntactic complexity measure and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492-518.
Ortega, L., and Byrnes, H. 2008. *The longitudinal study of advanced L2 capacities*. London: Routledge
Pekarek-Doehler, S. and Berger, E. 2016. L2 interactional competence as increased ability for context: A lomgitiudinal study of story opening. *Applied Linguistics*, online version
Pallotti, G. 2009. CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.
Plonsky, L. and Oswald, F. L. 2014. How Big Is "Big"? Interpreting Effect Sizes in L2 Research. *Language Learning, 64*(4), 878–912.
Polat, B., and Kim, Y. 2014. Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied linguistics, 35*(2), 184-207.
Ranta, L., & Meckelborg, A. 2013. *How much exposure to English do international graduate students really get? Measuring language use in a naturalistic setting. Canadian Modern Language Review, 69, 1–33.*
Revesz, A., Ekiert, M. & Torgersen, E. 2016. The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance. *Applied Linguistics, 37*(6), 828-848.
Robinson, P. 2011. *Second language task complexity: researching the cognition hypothesis of language learning and performance* (Vol. 2). Amsterdam: John Benjamins.
Robinson, P. 2015. The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 87-121). Amsterdam: John Benjamins.
Saito, K. 2015. Experience effects on the development of late second language learners' oral proficiency. *Language Learning*, *65*, 563-595.
Schegloff, E. 2001. Accounts of conduct in interaction: Interruption, overlap, and turn-taking. In J. Turner (Ed.), *Handbook of sociological theory* (pp. 287-321). New York: Kluwer Academic.
Schiffrin, D. 2001. *Discourse markers: Language, meaning, and context*. London: Blackwell.
Schmid, M. S., and Fägersten, K. B. (2010). Disfluency markers in L1 attrition. *Language learning, 60*(4), 753-791.
Shardakova, M. 2005. Intercultural pragmatics in the speech of American L2 learners of Russian: Apologies offered by Americans in Russian. *Intercultural Pragmatics, 2*(4), 423-451.
Skehan, P. 2001. Tasks and language performance assessment. In M. Bygate, P. Skehan and M. Swains (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 167-185). Harlow, Essex: Longman.
Skehan, P. 2003. Task-based instruction. *Language Teaching, 36*, 1-14.
Skehan, P. 2009. Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510-532.
Skehan, P. 2014. The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 1-26). Amsterdam: John Benjamins.

Skehan, P. 2015. Limited Attention Capacity and Cognition. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 123-155). Amsterdam: John Benjamins Publishing.
Smith, M. S., and Truscott, J. 2005. Stages or continua in second language acquisition: A MOGUL solution. *Applied Linguistics, 26*(2), 219-240.
Tavakoli, P. (2016) Speech fluency in monologic and dialogic task performance. *IRAL (*Special Issue: New directions in L2 speech fluency). *54(2)*: 133-151
Tavakoli, P. 2016. Speech fluency in monologic and dialogic task performance. *IRAL,* Special Issue on fluency. 54(2): 131-150.
Tavakoli, P. & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning. 58*(2): 439-473.
Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language (239-277).* Amsterdam: Benjamins.
Tavakoli, P., Campbell, C. & McCormack, J. (2016). Development of speech fluency over a short period of time: Effects of pedagogic Intervention. *TESOL Quarterly*. 50(2): 447-471.
TEEP (2014). Test of English for Educational Purposes. Reading: University of Reading. https://www.reading.ac.uk/web/FILES/ISLC/TEEP_candidates_handbook.pdf
Tidball, F., and Treffers-Daller, J. 2008. Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French language studies, 18*(03), 299-313.
Vercellotti, M. L. 2015. The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, 1-23.
Verspoor, M., De Bot, K., and Lowie, W. 2011. *A dynamic approach to second language development: Methods and techniques* (Vol. 29). Amsterdam: John Benjamins Publishing.
Walsh, S. 2013. Classroom discourse and teacher development. Edinburgh: Edinburgh University Press.
Warga, M., and Schölmberger, U. 2007. The acquisition of French apologetic behavior in a study abroad context. *Intercultural Pragmatics, 4*(2), 221-251.
Witton-Davies, G. 2014. The study of fluency and its development in monologue and dialogue. *Unpublished doctoral dissertation). University of Lancaster, Lancaster, England*.
Wray, A. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
Young, R. 2003. Learning to talk the talk and walk the walk: Interactional competence in academic spoken English. *North Eastern Illinois University Working Papers in Linguistics*, 2, 26-44.
Young. R. 2011. Interactional competence in language learning, teaching, and testing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 426-443). London: Routledge.

Table 1: Descriptive statistics indicating L2 development in CALF and DMs

	Monologue			**Dialogue**		
Measures	***Time 1 monologue***	***Time 2 monologue***	***Gains***	***Time 1 dialogue***	***Time 2 dialogue***	***Gains***
Length of AS unit	11.86 (2.22)	12.27 (2.16)	.42 (.24)	15.42 (3.35)	16.90 (3.03)	1.47 (3.65)
Ratio of	1.65	1.74	.09	2.18	2.46	.28

subordination	(.30)	(.34)	(.39)	(.36)	(1.20)	(.44)
Length of clause	6.89 (.91)	7.22 (1.18)	.36 (1.39)	6.75 (.98)	6.91 (1.20)	.16 (1.28)
WCR	.68 (.11)	.74 (.10)	.06 (.11)	.66 (.12)	.69 (.13)	.03 (.11)
% correct verbs	80.15 (14.41)	81.49 (14.11)	1.34 (18.57)	76.94 (10.74)	79.32 (9.67)	2.37 (5.58)
Length of run	6.68 (2.57)	8.17 (3.65)	1.61 (3.31)	7.94 (2.74)	9.27 (3.35)	1.33 (2)
Speech rate	143.35 (23.29)	160.97 (25.29)	19.16 (25.67)	175.23 (22.19)	192.32 (28.29)	17.09 (27.46)
Clause-internal pauses	13.51 (5.36)	13.67 (5.86)	.03 (4.81)	13.46 (6.67)	11.86 (7.30)	-1.60 (7.34)
Clause-external pauses	8.81 (3.16)	8.37 (2.92)	-.51 (3.11)	8.38 (3.46)	8.63 (3.80)	.25 (4)
Repair measures	7.27 (3.52)	6.62 (3.36)	.85 (4.11)	5.48 (4.11)	4.68 (3.31)	.80 (2.27)
D	45.38 (10.82)	46.42 (10.62)	.57 (15.78)	40.01 (11.09)	47.11 (11.52)	7.10 (13.86)
MTLD	35.44 (12.39)	38.28 (12.80)	2.08 (17.99)	27.89 (8.15)	36.58 (9.94)	8.68 (12.43)
Total number of DMs	4.41 (1.72)	5.81 (1.85)	1.41 (2.86)	9.64 (4.09)	10.52 (3.88)	.88 (4.94)
2-word DMs	.43 (.68)	1.27 (1.01)	.83 (1.30)	2.62 (1.59)	3.87 (1.90	1.16 (2.31)
3-word and longer DMs	.08 (.36)	.43 (.60)	.36 (.68)	.58 (.76)	1.63 (1.33)	1.05 (1.39)

n= 37 in monologues and *n*=35 in dialogues

Table 2: Number of discourse markers across tasks and times

Discourse markers	**Monologues**			
	Time 1	***Examples***	***Time 2***	***Examples***
One-word	144	and, but, so, actually, first, second, last	152	and, but, so, first, second, actually, finally
Two- word	16	I think, and then, after that	47	I think, and then, after that
Three-word & longer	3	First of all	16	In the beginning, as you know, at that time,
Total	163		215	
	Dialogues			

One-word	232	and, but, so, actually, OK, eventually	188	and, but, so, actually, OK, eventually
Two-word	99	and then, but actually, I think,	142	and then, you know, but actually, I mean, kind of
Three-word & longer	22	In the beginning, first of all, as I said	60	first of all, the first time, I know what you mean, I see where you are coming from, sorry for interrupting you, actually you're right
Total	347		390	

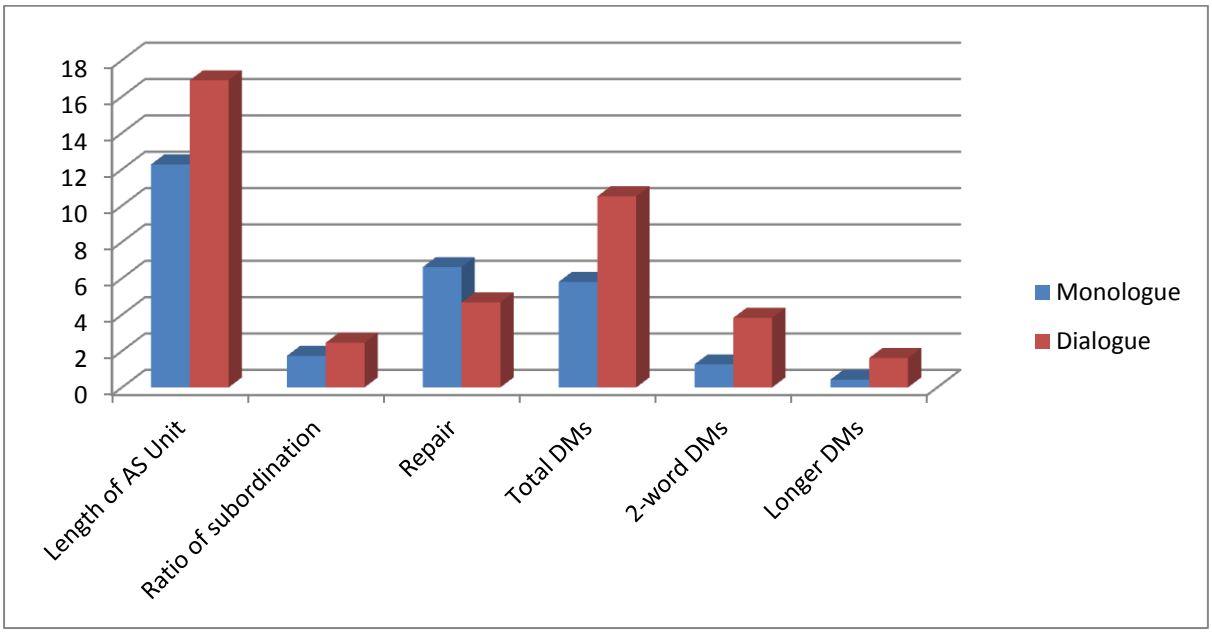

Figure 1: Significant differences between monologic and dialogic performance